SCIENTIFIC REPORT - NetWordS - Short Visit Grant 4730
Marie-Aude Lefer (Université catholique de Louvain, Belgium)
Research partner: Bruno Cartoni (Université de Genève, Switzerland)

Negative affixes in English, French and Italian:

A corpus-based contrastive approach

1. Purpose of the visit

The aim of our project, which is situated in the under-researched field of contrastive word-formation (Lefer & Cartoni 2011, Lefer 2011), is to provide a cross-linguistic description of the use of negative affixation in three European languages, namely English, French and Italian (i.e. the English *de-*, *dis-*, *in-*, *non-*, *un-* and *-less* and their French and Italian equivalents). We will achieve this goal by using trilingual corpus data, and more specifically translation corpus data.

Contrastive word-formation studies are relatively few and far between and, interestingly, they very rarely rely on corpus data. In this project, which takes stock of recent advances in corpus-based contrastive linguistics, the *Europarl* translation corpus (Koehn 2005) is used with a view to uncovering the main similarities and differences between negative affixes in the three languages investigated. The English-to-French and English-to-Italian translation data extracted from a 9-million-word corpus have been broadly examined in Cartoni & Lefer (2011), bringing to light many interesting cross-linguistic contrasts that are still largely unaccounted for at this stage. It appears, for example, that only a minority of English *un*-X-*ed* adjectives, where X is the verbal base of the derivative, have prefixed equivalents in French and Italian. The translation data also make it possible to unearth the major devices used to convey negation in each language (not only morphologically but lexically and syntactically as well).

In addition to its descriptive objectives, the project ultimately aims at feeding an existing online multilingual database of word-forming elements, Mulexfor, which is available in the form of a prototype.[1]

The aim of our short visit (which lasted 5 days) was two-fold: (1) analysing in more detail the *Europarl* English-to-French and English-to-Italian data sets; (2) implementing Cartoni & Lefer's (2011) results in the Mulexfor multilingual database of word-forming elements.

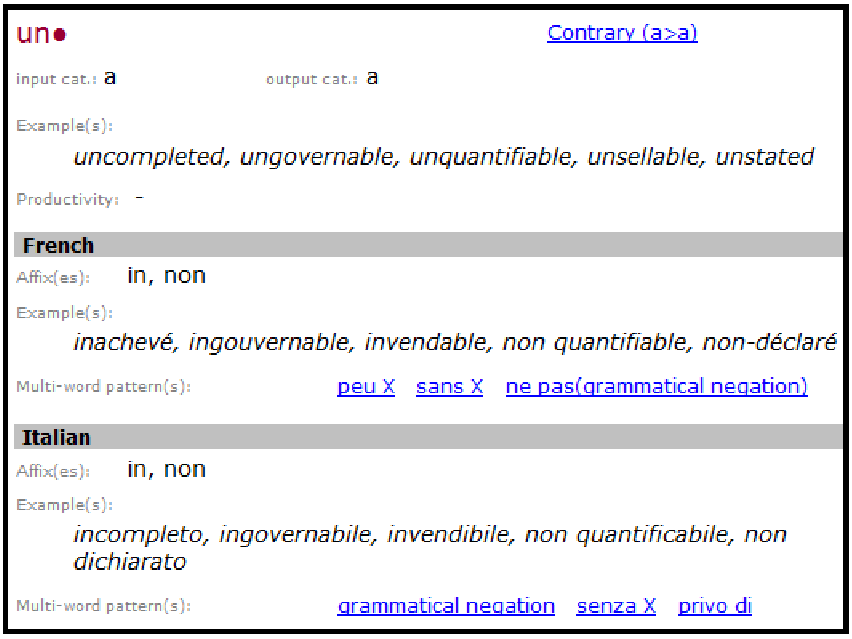
2. Description of the work carried out during the visit

The following work was carried out during the visit:

- **Implementation of Cartoni & Lefer's (2011) results in the Mulexfor database**

  - The findings described in Cartoni & Lefer (2011) have been implemented in the Mulexfor database. Concretely, trilingual corpus data was used to inform the content of the entries of the English negative affixes *de-*, *dis-*, *in-*, *un-* and *-less*. This implementation step required the semantic annotation of the data (using the following semantic categories: contradictory and contrary negation, privation, reversal and removal) as well as the semantic disambiguation of the *un-* and *dis-* data sets. Two types of French and Italian equivalents were inserted in the trilingual entries: (i) morphological equivalents of English negative affixes and (ii) multiword patterns and paraphrases that are frequently used in French and Italian to render English negative affixes (so-called 'non-morphological equivalents') (e.g. EN *it resulted from the inexperience of the German authorities* vs FR *c'était dû au manque d'expérience des autorités allemandes*). This is illustrated in Figure 1 below. We also added contextualised examples taken from Europarl, as shown in Figure 2.

[1] https://sites.google.com/site/mulexfor/



un● Contrary (a>a)

input cat.: a output cat.: a

Example(s):

uncompleted, ungovernable, unquantifiable, unsellable, unstated

Productivity: -

French

Affix(es): in, non

Example(s):

inachevé, ingouvernable, invendable, non quantifiable, non-déclaré

Multi-word pattern(s): peu X sans X ne pas(grammatical negation)

Italian

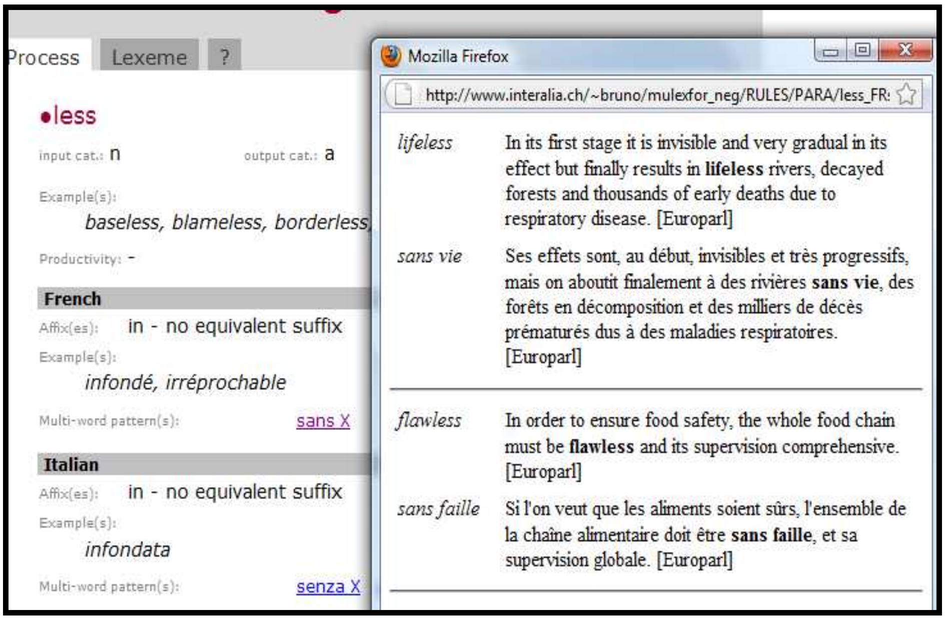
Affix(es): in, non

Example(s):

incompleto, ingovernabile, invendibile, non quantificabile, non dichiarato

Multi-word pattern(s): grammatical negation senza X privo di

Figure 1: Corpus-informed *un*-entry in Mulexfor



●less

input cat.: n output cat.: a

Example(s):

baseless, blameless, borderless,

Productivity: -

French

Affix(es): in - no equivalent suffix

Example(s):

infondé, irréprochable

Multi-word pattern(s): sans X

Italian

Affix(es): in - no equivalent suffix

Example(s):

infondata

Multi-word pattern(s): senza X

http://www.interalia.ch/~bruno/mulexfor_neg/RULES/PARA/less_FR:

lifeless	In its first stage it is invisible and very gradual in its effect but finally results in **lifeless** rivers, decayed forests and thousands of early deaths due to respiratory disease. [Europarl]
sans vie	Ses effets sont, au début, invisibles et très progressifs, mais on aboutit finalement à des rivières **sans vie**, des forêts en décomposition et des milliers de décès prématurés dus à des maladies respiratoires. [Europarl]
flawless	In order to ensure food safety, the whole food chain must be **flawless** and its supervision comprehensive. [Europarl]
sans faille	Si l'on veut que les aliments soient sûrs, l'ensemble de la chaîne alimentaire doit être **sans faille**, et sa supervision globale. [Europarl]

Figure 2: Contextualised examples in Mulexfor: the case of -*less*

- **Analysis of *un*-derivatives with 3 or more occurrences and their French equivalents**

  - In Cartoni & Lefer (2011), which is based on the analysis of *hapaxes legomena* (1 occurrence in the corpus) and *hapaxes dislegomena* (2 occurrences), we assumed that (i) in order to investigate the translation patterns of affixes in an electronic corpus and uncover cross-linguistic contrasts, it is advisable to focus on *hapaxes (dis)legomena* and (ii) high-frequency derivatives tend to be translated into a very limited number of equivalents in the target languages, and are consequently of little morphological interest. During the short visit, we wanted to test whether there was indeed a link between the frequency of derivatives and the number of different equivalents they correspond to in the target languages. In order to test this, we analysed all the *un*-words (whether frequent or not) in a 3-million-word *Europarl* subcorpus (1996-1999), where English is clearly identified as the source language and French and Italian as target languages (see Cartoni & Meyer 2012). We also categorised the French equivalents of the English *un*-tokens (using the following three categories: prefixed word, simplex word and paraphrase). A total of 2059 *un*-tokens and their French translation equivalents were analysed.

- ❖ **Extraction and preliminary analysis of *un*-derivatives in English target texts translated from French, Italian (and German)**

  - ➢ In a second step of our work, we extracted and performed a preliminary analysis of the *un*-words contained in English target texts translated from French, Italian and German (each subcorpus contains ca 2 million words in total). Our ultimate objective here is two-fold: (1) determine whether the use of English *un*-words is different in translated English from original English, (2) find out whether the source language (French, Italian or German) plays a role on the use of *un*-words in English translated texts. More particularly, we would like to test whether English *un*-X-*ed* words, which have been shown to rarely have prefixed equivalents in French and Italian (Cartoni & Lefer 2011), are under-used in English texts translated from these two languages (cf. Tirkkonen-Condit's 2004 'Unique Items' Hypothesis).

3. Description of the main results obtained

- ❖ Five trilingual affix entries have been enriched with the help of corpus data. The implementation step in **Mulexfor** showed that translation corpora can help lexicographers:

  - ➢ Identify correspondences between derivational affixes and their target language morphological equivalents;
  - ➢ Uncover systematic paraphrase patterns in cases of marked cross-linguistic differences;
  - ➢ Select authentic examples.

- ❖ Study on the ***un*-words in original English and their equivalents in translated French**

  - ➢ The results indicate that high-frequency *un*-derivatives in Europarl (e.g. *unable*) display a large number of possible equivalents in French (20 different equivalents out of 77 occurrences of *unable*, such as *incapable*, *ne pas pouvoir*, *ne pas être en mesure*, *ne pas savoir*, *ne pas avoir la possibilité de*). Similar examples include: *unfair* (51 tokens, 10 French translation equivalents), *uncertainty* (38 tokens, 11 equivalents), *undoubtedly* (28 tokens, 12 equivalents), *undemocratic* (14 tokens, 6 equivalents), *unfamiliar* (8 tokens, 7 equivalents), etc. It might therefore be concluded that contrastive word-formation studies based on translation corpora should not be restricted to *hapaxes* (*dis*)*legomena*, as this can obscure important translation patterns. It still needs to be determined whether these higher-frequency derivatives and their equivalents can be used to shed some new light on cross-linguistic morphological contrasts. In any case, they undoubtedly help understand translation-related lexical phenomena such as the diverging polysemy and the partial phraseological equivalence of derivatives cross-linguistically (see Lefer in preparation).

- ❖ Study on the ***un*-words in translated English**

  - ➢ The data show that *un*-words are more frequent in English target texts translated from Italian (198 *un*-tokens per 100,000 words), French (197 *un*-tokens per 100,000 words) and German (179 *un*-tokens per 100,000 words) than in original English texts (149 *un*-tokens per 100,000 words). Interestingly, we observe a (slightly) lower type-token ratio in translated English (0.13 on average) than in original English (0.16), indicating that translators resort to a rather limited range of different *un*-lexemes compared to original English. This trend ties in with research on translation universals, which has shown that translated language tends to be lexically poorer than non-translated, original language ('lexical simplification'; see Baker 1995).

4. Future collaboration with host institution

The short visit has made it possible to extract new data sets and start analyzing them. The following analyses now need to be carried out:

- ***Un*-words in original English and their equivalents in translated French & Italian**
    - Analyse the Italian translations and draw general conclusions on the basis of the full data set;
    - Compare the equivalents of the high-frequency *un*-words in the corpus and in English-French (and English-Italian) bilingual dictionaries. If the majority of equivalents are listed in dictionaries, it would probably be advisable to use lexicographic data rather than corpus data to examine the translation patterns of these derivatives.

- ***Un*-words in translated English**
    - Analyse the items in French and Italian source texts that led to the use of *un*-words in translated English (prefixed words, non-prefixed words, multiword patterns, etc.) to refine the description of the cross-linguistic overlap of English, French and Italian negative affixes sketched in Cartoni & Lefer (2011);
    - Investigate the *un*-X-*ed* pattern, which is expected to be less frequent in English translated from French or Italian, as there are very few corresponding prefixed words in these two languages. English data translated from original German texts will also be examined (German, like English, is a Germanic language, whereas French and Italian and Romance languages).

These further analyses will be carried out in collaboration with Bruno Cartoni (University of Geneva).

5. Projected publications/articles resulting or to result from your grant

We plan two publications in scientific journals (probably in early 2013):

- *Corpus Linguistics and Linguistic Theory* (methodological and theoretical aspects central to the analysis of prefixed words in translation corpora, e.g. influence of the frequency threshold and the corresponding translation patterns);

- *Meta* (study of *un*-words in original and translated English, impact of the source language on the use of derivatives in target texts, under-use of *un*-X-*ed* derivatives in English translated from French, Italian and/or German).

References

Baker M. (1995) Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7(2). 223-242.

Cartoni B. & M.-A. Lefer (2011) Negation and lexical morphology across languages: insights from a trilingual translation corpus. *Poznan Studies in Contemporary Linguistics*, 47(4). 795-843

Cartoni B. & T. Meyer (2012) Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. Proceedings of LREC 2012, May 23-25 2012, Istanbul, Turkey.

Koehn P. (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Lefer M.-A. (in preparation) Word-formation in translated language: the impact of language-pair specific features and genre variation.

Lefer M.-A. (2011) Contrastive word-formation today: Retrospect and prospect. *Poznan Studies in Contemporary Linguistics*, 47(4). 645-682.

Lefer M-A. & B. Cartoni (2011) Prefixes in contrast. Towards a meaning-based contrastive methodology for lexical morphology. *Languages in Contrast* 11(1). 86-104.

Tirkkonen-Condit S. (2004) Unique items - over- or under-represented in translated language? In: Mauranen, A. & Kujamäki, P. (eds) *Translation Universals. Do they exist?* Amsterdam & Philadelphia: Benjamins. 177-184.